DOCUMENT RESUME

ED 227 649 FL 013 374

AUTHOR Ingram, D. E.
TITLE New Approaches to Language Testing.
PUB DATE Sep 82
NOTE 42p.; Paper presented at the Annual National Languages Conference of the Australian Federation of Modern Language Teachers Associations (4th, Perth, Australia, September 3-6, 1982).
PUB TYPE Speeches/Conference Papers (150) -- Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Adolescents; Adults; Comparative Analysis; English (Second Language); Evaluation Methods; Foreign Countries; *Language Proficiency; *Language Tests; Secondary Education; Second Language Learning; Second Language Programs; Student Evaluation; Testing; Test Reliability; Test Validity
IDENTIFIERS Australia; *Australian Second Language Proficiency Ratings; Foreign Service Institute Interview

ABSTRACT

The role of different types of testing instruments in second language programs is discussed. Knowledge-based tests, including discrete-point and integrative tests are contrasted with direct tests, including the Foreign Service Institute Scale and the Australian Second Language Proficiency Ratings (ASLPR). The ASLPR is described in detail in terms of its development, levels, tasks, parameters, uses, and formal trials. The ASLPR was originally developed to measure the proficiency of adult English as second language learners and then expanded for use with adolescents and other languages. The ASLPR is valid, reliable, and suitable for the measurement of general proficiency, while other instruments may be more appropriate for various other testing objectives. Other uses for language proficiency tests in second language programs include obtaining a quick group measure of proficiency, measuring special purpose proficiency; measuring summative evaluation of course attainment; and for student placement, diagnosis of student strengths and weaknesses, program evaluation, and test validation. (RW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED227649

NEW APPROACHES TO LANGUAGE TESTING

D.E. INGRAM

Department of Languages and Literature,
Brisbane College of Advanced Education,
Mount Gravatt Campus, Brisbane, Queensland.

Paper to the Fourth National Languages Conference of the Australian Federation of Modern Language Teachers Associations, The Language Curriculum in the 1980's - Affective, Effective or Defective, Perth, 3rd - 6th September, 1982.

FL 013 374

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D. E. Ingram

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

C O N T E N T S

Abstract

I Introduction

II Knowledge-based Tests

II.1 Discrete-Point Tests

II.2 Integrative Tests

II.3 Knowledge versus Proficiency

III Direct Tests

III.1 FSI Scale

III.2 Development of the ASLPR

III.3 Description of the ASLPR

- Number of levels
- Describing overall language behaviour
- Tasks and how carried out
- General proficiency vs. register proficiency
- Underlying proficiency vs. its realisation in a task
- Parameters of change
  - range of situations
  - complexification and personal meaning
  - register
  - beyond the surface
- The three columns
- Naming the proficiency levels
  - description
  - number
  - title

III.4 Rating Learners' Proficiency

III.5 Uses being made of the ASLPR

- Integrating framework for programme planning
- Streaming
- Professional registration of teachers
- Selection for ESP courses
- Court of law
- Sociolinguistic surveys
- Secondary School FL assessment
- Programme evaluation
- Selection of foreign students

III.6 Formal Trials

III.6.1 Validity

III.6.2 Reliability

III.6.3 Trial Conclusions

IV Tests in Language Programmes

1. To measure how well a learner can use the language, i.e. to measure proficiency
2. To give a quick group measure of proficiency
3. To measure "special purpose" proficiency
4. To measure proficiency in a "graded objectives" course
5. To measure course attainment (summative purposes)
6. To show how well the learners are mastering what is being taught (formative purposes)
7. To stream learners into an appropriate course
8. To diagnose learners' strengths and weaknesses
9. To evaluate a second language programme
10. To validate tests

V Conclusion

-oOo-

ABSTRACT

Recent developments in language *teaching* methods that have emphasized communicative skills have stimulated the development of new approaches to language *testing*. These have been paralleled in general education by interest in competency-based syllabuses and assessment. One outcome of these movements has been work around the world on the development of proficiency rating scales. It is generally recognized in the world-wide language testing field that Australia has been in the forefront of this with the emergence of the Australian Second Language Proficiency Ratings (ASLPR), developed by Ingram and Wylie originally for the Federal Department of Immigration and Ethnic Affairs. This instrument has now been re-considered and applied to Secondary School foreign language teaching and is reflected in the approach to assessment adopted in the new functional syllabuses being introduced to Queensland Secondary Schools. Yet direct instruments such as the ASLPR are not the only appropriate test instruments and teachers need to see the ASLPR and direct instruments generally in their proper perspective in a total teaching-testing programme. This paper will refer to the place of the ASLPR in foreign and second language teaching, indicate its appropriate place, and discuss and illustrate the place of more traditional ("indirect" and "semi-direct") instruments in a teaching-testing programme. Thus the paper will seek to provide a comprehensive and balanced perspective on language testing in a modern, language teaching programme, identifying the aims of testing at different times and the most appropriate instruments to achieve the aims.

-oOo-

I INTRODUCTION:

Few professions experience such rapid change, in philosophy if not in practice, as modern language teaching. In twenty-five years we have passed from grammar-translation approaches through audiolingualism and audiovisualism, situational methods, and cognitive code to functions, notions, communicative approaches and graded objectives. On the way, we have encountered community language learning, community involvement, immersion classes, compact, intensive and "special purposes" courses, individualization, self-directed learning, self-access, and many others. The experienced language teacher's anthem could well be "I've been everywhere, man."

However and, perhaps, fortunately, the apparent conservatism of practising language teachers and the stolidity of the school systems within which they work have provided some counter-balance to this plethora of change so that the flux of philosophies or methods is not reflected in such rapid change in classroom practice, at least at the systemic level. While such stability may prevent costly waste as new courses are developed, adopted, and discarded, it becomes stolidity when well-balanced innovations are ignored or when changes in part of the language teaching system (e.g. goals, methods, or course books) are not reflected in matching change in other parts (e.g. in testing or examinations).

II KNOWLEDGE-BASED TESTS:

II.1 Discrete-Point Tests

In principle at least, changes in one's understanding of the nature of language and of how it is learned lead to changes in goals, syllabus, methods, and, our concern here, testing. While language was seen as patterns learned by stimulus-response, as described in behaviourist psychology, one way to test was to provide a stimulus and see whether the learner could provide the appropriate response; so test items commonly took forms such as this:

Tick the item that best fills the gap

Question: "How old is Joe?"
"He's two years older ____ his sister Mary."

Answer: that ☐ as ☐ than ☐ of ☐

This type of item was also appropriate if one saw language in more cognitivist, rule-based terms, i.e., one could say that the item is testing whether the learner knows and can apply the rule that you use the conjunction *than* after a comparative. In fact, whether one describes knowledge in behaviourist or cognitivist terms, this type of item is basically a test of knowledge, specifically a test of knowledge of a discrete point of grammar, and so this approach to language testing has been called *discrete-point testing*.

II.2 Integrative Tests

One problem with discrete-point testing is that it isolates the language components whereas, when language is being used in real life, all the components are together, supporting each other in meaning and dependent on each other structurally; part of the skill of using language involves being able to put all the components together and to comprehend them when they are received together. Several approaches to testing have sought, therefore, to integrate the language components into a total language event and to test knowledge of them as discrete items in the context of that event. Such test types can be called *integrative* tests. Depending on how they are marked, translation and guided compositions could be seen as integrative tests. So in a translation test such as the following, interest focussed on whether the learner knew the Motu equivalents for each vocabulary item and whether he knew such grammatical rules as that, in Motu, the verb goes at the end of the sentence and prepositions follow the nouns they govern, e.g.

Translate the following into Motu:

"Look at that man."
"He is sitting near the door.
Who is he?"
"He is my father."

⟹ "Una tau a itaia."
"Tau iduara badinai helai.
Daika tau?"
"Una lau tamagu."

This type of test was characteristic of traditional grammar-translation methods even though, at that time, it might not have been called an integrative test, and, in fact, was often marked more like a discrete-point test containing a multitude of separate items. Such tests have many problems, three of which are these: first, they often do not sample the learner's knowledge very comprehensively since the person writing or selecting the passage for translation often tries to include things that he sees as learning problems with the result that, as a test of the learner's knowledge, it becomes unbalanced; second, translations, especially if they are "difficult", tend to be carried out by the learner on an isolated word-by-word basis, which not only leads to linguistic oddities but destroys the exercise as an integrative test; third, once the language becomes more sophisticated than the example used above, marking becomes both onerous and subjective with the marker having to decide whether to accept or how much to penalize the various translations given for each discrete lexical or syntactic item. Consequently, other types of integrative tests became popular in the 'sixties and 'seventies, tests such as cloze, dictation or white noise. When dictation is given properly [cf. Oller 1979:267-282], each dictated "chunk" is long enough to exceed the learner's capacity to retain the words in immediate memory and so the forgotten items have to be filled in from the context and the learner's knowledge of the language. With cloze and white noise tests, items are deleted or masked by noise and again the learner has to draw on the context and his knowledge of the language to identify what is missing. A typical cloze passage of the desirable 40 to 50 deletion length could be as follows:

Cloze Test

Name: _____ Date: _____

Native Language: _____ Class: _____

Dialect: _____ Year of Course: _____

Replace the words that have been deleted in the following text.
Insert one word only in each space marked thus _____.

It was the lunch hour and Brandon went across to the grand old public library. The huge reading rooms with their tapestried walls and the enormously high ceilings were nowhere near as crowded as usual. A few scattered students sat at the long __(1)__, groups of tourists strolled about looking at the __(2)__ and the cil paintings, speaking in hushed whispers __(3)__ they moved from room to room. As Brandon __(4)__ through the marble hallways towards the complex of __(5)__ in the west wing, he remembered the months __(6)__ had spent in these rooms awakening his mind __(7)__ a world he had known so little about. __(8)__ had not exaggerated when he had spoken to __(9)__ students: it was here, under the devoted guidance __(10)__ one man, that he had learned more about __(11)__ world of science than he had, later on __(12)__ he started to study at Oxford. This library __(13)__ his finest school, the man he was about __(14)__ see after so many years his most accomplished __(15)__. He wondered whether the school or his teacher __(16)__ help him now. Somewhere in the thousands of __(17)__ that outlined the history of the pre-war years, __(18)__ information that he sought must be located. Would __(19)__ old teacher know where to look?

He turned __(20)__ into the corridor lined with glass-panelled office doors, __(21)__ dark except one at the end of the __(22)__ where a dim light was on inside. The __(23)__ was blocked intermittently by the silhouette of a __(24)__ passing back and forth in front of a __(25)__ lamp. It was old Doctor McNamara's office, the __(26)__ one he had occupied for over a quarter __(27)__ a century; the slowly moving figure beyond the __(28)__ was unmistakably that of the scholar.

Brandon walked __(29)__ to the door and knocked softly. The dark __(30)__ moved instantly behind the glass. The door opened __(31)__ Doctor McNamara stood there, his wrinkled face smiling, __(32)__ eyes behind the thick lenses of his spectacles __(33)__ as he peered at his former student. "Come __(34)__," he said. "What is the help you need?"

"__(35)__ is a long story", said Brandon, "but I __(36)__ tell you everything I can. I must, because __(37)__ need your help".

The better part of half __(38)__ hour passed as Brandon told the old scholar __(39)__ he needed to know and why it was __(40)__

urgent for him. McNamara listened in silence until ___(41)___ had finished speaking. Then he rose and led ___(42)___ out to a far corner of the library. ___(43)___ took a book off the shelf, opened it ___(44)___ indicated to Brandon that he should read. The ___(45)___ leant over, scanning the words. The entry McNamara ___(46)___ indicated detailed the deaths of a Colonel William ___(47)___ and his family during an attack by rebel ___(48)___ one tragic afternoon in October 1924. It described ___(49)___ full details the final minutes of fighting as ___(50)___ Andrews struggled to repel the attacking mob, firing weapons from the windows and finally surrendering as their resources gave out. This was the evidence Brandon needed; the last clue to the puzzle lay in front of him and he could now proceed to complete his plans.

CLOZE TEST ANSWERS

1. table	11. the	21. all	31. and	41. Brandon
2. books	12. when	22. hallway	32. his	42. him
3. as	13. was	23. light	33. half-closed	43. He
4. walked	14. to	24. figure	34. in	44. and
5. offices	15. teacher	25. desk	35. It	45. latter
6. he	16. could	26. same	36. will	46. had
7. to	17. volumes	27. of	37. I	47. Andrews
8. He	18. the	28. glass	38. an	48. tribesmen
9. his	19. his	29. up	39. what	49. in
10. of	20. left	30. figure	40. so	50. the

II.3 Knowledge versus Proficiency

The types observed thus far have been called *discrete-point* and *integrative* tests. As described, they basically test for knowledge either in isolated discrete items or integrated into a language event where the learner has the inherent redundancy of language to assist him in recalling the item. Discrete point and integrative tests can, respectively, also be described as *indirect* and *semi-direct*, especially when they are used to estimate a learner's proficiency. Discrete points tests are *indirect* in that they assume that a measure of one thing, viz., knowledge or linguistic competence, can serve to measure something else (viz. language proficiency). Cloze, dictation or white noise tests are *semi-direct*, in that, they incorporate each item into a language event more closely involving all the sub-skills of the learner's total proficiency. Neither type

gives a direct measure of proficiency but they rank-order the learners and, from that rank-ordering, one can say which learners are more or less proficient than the others. In that such tests let you compare learners, they can also be called norm-referenced tests.

The problem when tests such as these are used to assess proficiency is that they test one thing, e.g., knowledge, in order to make a statement about something else, viz., proficiency. Yet knowledge is not the same as proficiency; a learner can have much knowledge about a language, be able to state many grammatical rules, and answer correctly many tests of vocabulary and yet not be able to mobilize that knowledge readily enough to carry out communication tasks effectively. In a study of the EFL proficiency in a foreign languages institute in China in December 1980, the present writer found the students scoring very high on the knowledge-based CELT Structure test yet their ability to carry out basic communication tasks as reflected in ratings on a proficiency rating scale was very low [Ingram 1980]. This contrast between a learner's knowledge and his ability to carry out real-life communication tasks is a familiar problem to foreign language teachers [cf. Angelis 1979; Oller and Spolsky 1979]. Consequently, as the teacher's concept of language has broadened from a body of knowledge to a means of carrying out communication tasks and as language teaching has shifted from a process of knowledge transmission to focus on communication goals using communicative methods to develop practical communication skills, so the need became urgent for the development of test instruments that focus on those communication skills and measure practical language proficiency. In Queensland, for example, competency-based assessment is being introduced across all Secondary School assessment. In foreign languages, "competency" can only refer to the ability to carry out communication tasks and so the need has been to find ways of measuring and stating that ability validly and reliably.

III DIRECT TESTS:

In contrast to semi-direct and indirect tests which, we have seen, rank order learners or measure one thing (especially knowledge) and try to make a statement about something else (e.g. language proficiency), *direct* tests focus on actual language behaviour, i.e., they focus directly on the learner's proficiency as demonstrated in the way he carries out actual communication tasks; proficiency statements are then related to and made in terms of the learner's actual language behaviour. Thus, rather than just measure knowledge, *direct* instruments describe different levels of proficiency in terms of the communication tasks that the learner can carry out and how they are carried out. Learners are rated by being matched against the level on a scale consisting of a series of proficiency descriptions that best describes their language behaviour. In other words, in direct testing, a learner's proficiency is assessed by being matched against a concept of how proficiency develops and such tests can, therefore, be called criterion-referenced tests.

III.1 FSI Scale

Few instruments of this sort have been developed. The one that has been in use longest and most widely has been the "Absolute Language Proficiency Ratings" of the United States Foreign Service Institute School of Language Studies [1968], generally known as the FSI Scale though recently re-named the ILR Scale, presumably, the International Language Rating Scale. However, this scale has severe limitations, especially when used in a school programme or in a large migrant programme with wide variations in the learners' educational backgrounds [see Ingram 1980a, 1981 where the limitations are delineated]. Not least, it fails to differentiate very closely amongst low-level learners, it was designed for use with well-educated civil and foreign service personnel, and it measures only conversation skills and reading. Thus, the FSI Scale is not very useful where the learners vary widely in educational and social background, where one wishes to differentiate the proficiency levels of learners in low-level classes, or where all four macroskills are of interest. In recent years, therefore, Elaine Wylie and the present writer have been seeking to develop a direct instrument that would overcome these deficiencies. This has become known as the Australian Second Language Proficiency Ratings (ASLPR) [see Appendix].

III.2 Development of the ASLPR

Work on the ASLPR commenced in September 1978 as a result of a decision by the Australian Department of Immigration and Ethnic Affairs to develop new English courses for the Adult Migrant Education Program. The following steps were taken:

1. Drawing on the FSI Scale, psycholinguistic studies of second language development, and the intuitions of many years experience of teaching English, French and Italian as second or foreign languages, Elaine Wylie and the present writer[1] sought to describe language behaviour at nine proficiency levels along the developmental path from zero to native-like.

2. The emerging scale was re-assessed and elaborated in face-to-face interviews with 50 adult learners of English as a Second Language (ESL), 40 adolescent learners of ESL, and over 100 adolescent and adult learners of French, Italian and Japanese as foreign languages.

3. Early versions of the scale have been released for use in the Adult Migrant Education Program since February 1979, thus providing feedback on its adequacy as a result of its practical implementation.

4. Formal trials have been (or are being) conducted to assess the validity and reliability of the ASLPR when used with adult ESL learners, adolescent ESL learners, and adolescent and adult foreign language learners of French, Italian and Japanese.

III.3 Description of the ASLPR

The ASLPR describes language behaviour at nine proficiency levels along the developmental path from zero to native-like. An additional three intermediate levels are available for use at 2+, 3+ and 4+ though they are not described. Each macroskill is described separately so that the ASLPR actually consists of four separate but conceptually related scales and a learner's proficiency is stated in a profile indicating proficiency in each macroskill separately, e.g., S:1+, L:2, R:1, W:1.

The descriptions provide an overall picture of the behaviour

observable at each proficiency level so that a learner's proficiency is rated as a global assessment is made of which scale description best matches the observed behaviour. The descriptions are not, therefore, a checklist since the complexity of language behaviour and its development mean that different aspects may develop at different rates. So, for instance, a Vietnamese learning English may seem to develop his phonological and syntactic skills at different rates compared with a Scandinavian. During trials of the ASLPR, a Chinese businesswoman showed exceptional ability to comprehend sums of money even though her listening skills, overall, were rated at L:0+. This global assessment of a learner's proficiency in a macroskill may seem to make the assessment process more subjective because the rater has, essentially, to use judgment in balancing different aspects of development against each other but it is necessitated by the sheer complexity of language and its development.

The scale describes language proficiency in terms of the tasks the learner can carry out and how they are carried out. The descriptions, therefore, include specific reference to such factors as communication tasks, range of situations, functions, pronunciation, vocabulary, syntax, discourse, and register sensitivity and flexibility. In applying the ASLPR, the rater must continually refer to the scale and deliberately match the described characteristics with the components of the language observed.

The ASLPR has been designed to measure general proficiency by matching observed language behaviour against global descriptions. The concept of general proficiency raises some important issues.

First, language occurs only in situations and if the proficiency descriptions are related to particular situations, then one could be accused of measuring only proficiency in those specific situations, i.e., one would not be measuring general proficiency but proficiency in specific registers. On the other hand, language varies from situation to situation; it varies according to who is using it, to whom and about what subject; it varies according to where it is being used, what mood the person is in, what the speaker's role is, and so on. In other words, it would seem as though one cannot speak of general proficiency so much as proficiency in a language in this

situation or that, in this register or that. Yet such a view seems, at least, to be counter-intuitive. If we say that X speaks Chinese or Y speaks French, we do not mean that X can only give a lecture on engineering in Chinese or that Y can only talk about neurosurgery. Rather, when we say that someone can speak a language, we mean that that person can speak the language in the sorts of situations people commonly encounter. That is, there are certain everyday situations in which we, as human beings living in a physical and social world, are necessarily involved: we all must, for example, obtain food and shelter, we must find our way about, or we must establish relationships with other people. General proficiency, then, refers to the ability to use language in these everyday, non-specialist situations. In addition, even though there are observable differences between registers, no register is entirely discrete or it would not be recognized as part of the language and ability to use those commonly occurring features (e.g., of phonology, syntax, lexis, discourse or functions) would seem to be entailed in the concept of "general proficiency" and to underlie the learner's register flexibility or his ability to cope with new situations and to modulate his language to them.

Second, because the ASLPR is measuring "general proficiency" rather than just proficiency in certain defined situations, one has to distinguish the underlying general proficiency from its realization in a particular situation or a particular task. Someone who has never used public transport, for instance, may not know what to do and may not use language entirely appropriately if he has to buy a bus or train ticket. A foreign language learner in a country without door-to-door milkmen may be unable to write a note appropriately to change a daily milk order but, if the learner is to be rated at W:1 on the ASLPR, you would expect him to be able to show in some other situation (e.g., in a note to a teacher to explain an absence) that he is able to perceive a relationship and the information to be conveyed, and to select from his language repertoire the minimum language forms needed to carry out the note-writing task. In other words, the ASLPR seeks to measure the underlying general proficiency rather than the fulfilment of an absolutely specified task in an absolutely specified situation. Hence, in the ASLPR there is a "General Description" column that describes the language behaviour observable at a particular level and an "Example of Specific Tasks"

column that is not absolute nor a checklist but only exemplifies the observable behaviour. This also explains why the ASLPR is not accompanied by a set of standardized test items but, rather, the interviewer is required to vary the situations and tasks within which the learner is asked to demonstrate his proficiency so as to gain a picture of his underlying proficiency through his performance in a variety of situations and tasks.

This concept of general proficiency and how it permeates the use of language in all situations and tasks underlies the ASLPR, its behavioural descriptions, and the parameters of change through the scale. Though each description seeks to provide a fairly comprehensive picture of the learner's language behaviour at that level in that macroskill, different parameters of change are dominant at different stages of development. In addition, language development occurs in response to need and this basic motivational fact also influences the path language development takes and hence the dominant changes that occur through the ASLPR. At the lowest levels of the scale, the learner's language development reflects his immediate survival needs as determined, for the second language learner, by his need to survive in the L2 community or, for the foreign language learner, by the situations in which the teacher or course designer choose to place him. Thus, at the lowest proficiency levels, the dominant parameter of change is the range of situations in which the learner can use the language; at first he will use largely formulaic language or rote-memorized items with limited creativity and be able to carry out only those simple functions in those situations regularly encountered or specifically taught. Once immediate survival needs are met, the learner will be able to reach out to other people, to establish social relations and so, in the middle levels from 1+ on, the language complexifies through the emergence of modification devices, discourse development, embedding and alternate forms to enable the more precise statement and comprehension of personal meaning. As the learner's language base increases, so he can choose more deliberately from his repertoire to modify his language to match different situations and to extrapolate meaning for new items. Hence, by level 3, register sensitivity and flexibility become dominant distinguishing features. Beyond this level, especially from 4 to 5, the ability to see beyond surface meaning, to comprehend and use humour and irony, and to draw on significant cultural understanding becomes significant.

The ASLPR is laid out in three columns. The "General Description" column describes in some detail the language behaviour at each level and applies equally to all languages and to both foreign and second language learners. The middle column, "Examples of Specific Tasks", exemplifies the language behaviour observable and especially the sorts of tasks that the learner can carry out. This column differs slightly from language to language and from second to foreign language learners. The third column, the "Comment" column mainly explains or defines distinctive terms used in the other columns.

The proficiency levels are identified in three ways, first, by the elaborated description, second, by a number preceded by the macroskill initial (e.g., S:1+, L:2, R:1, W:1-), and, third, by a short descriptive title. The slightly peculiar numbering system was originally used for practical reasons during development of the scale but it now is unlikely to be changed principally because the whole number levels correspond to the whole number levels on the FSI Scale, thus facilitating cross-referencing. It has to be stressed that there are 9 defined plus 3 undefined levels at 0, 0+, 1-, 1, 1+, 2, 2+, 3, 3+, 4, 4+ and 5. Other intermediate levels should not be used since, first, they imply, in some cases such as the use of 0- for illiterate learners, a serious misunderstanding of the nature and process of proficiency measurement; and second, the proliferation of undefined levels further sub-dividing spans of proficiency is likely to lead to marked deterioration in reliability.

III.4 Rating Learners' Proficiency

It was observed earlier that the traditional concept of standardized test items is entirely inappropriate to the ASLPR as a measure or general proficiency. Rather, to rate a learner's language proficiency, his maximum language behaviour is elicited in whatever situations and tasks are most appropriate to him, it is observed, and matched against the scale descriptions. Although, as we observed earlier, this approach is subjective, it is not impressionistic and raters are required to refer to the scale continually and to carefully compare the learner's language behaviour with the scale descriptions. In order to elicit that behaviour, an interview process is used. The interview should be so structured that the language

activity the learner engages in is as natural as possible. For the learner, the interview should be as relaxed and informal as is appropriate to elicit his best language performance but, for the interviewer, it should be deliberately developed to explore all those features delineated in the scale that distinguish one level from another and to ensure that the full extent of the learner's language proficiency is observed and rated. Broadly, the interview has three parts: first, an *exploratory* stage in which the interviewer settles the learner down and gets an approximate idea of where the learner's proficiency falls and what topics it would be useful to pursue; second, an *analytical* stage in which the interviewer explores all the features of the learner's language behaviour, gives him an opportunity to use whatever language he can and leads him on to his "linguistic breaking-point"; and, third, a *concluding* stage in which the interview is rounded off and, for a moment or two, the interviewer reverts back to activities well within the learner's ability so as to avoid sending him away with a sense of failure after having been taken to his "linguistic breaking-point".

III.5 Uses being made of the ASLPR

Basically, the ASLPR does two things: first, it provides an overall picture of how a second language develops from zero to native-like and, second, it provides a means of measuring language proficiency in terms of practical language skills. Hence it can be used in any circumstances where these two things are relevant. Some of the actual uses to which the ASLPR has already been put include the following:

- as an integrating framework within which a language programme or series of language programmes can be coherently planned, e.g., the Adult Migrant Education Program in Australia;
- to stream ESL learners into migrant classes on a basis of their proficiency;
- to stream foreign language learners of Greek, Italian and other languages into appropriate College language classes;
- to assess whether overseas-trained applicants for teacher registration in Queensland have sufficient English proficiency to teach in Queensland schools;

- to select applicants for "English for Special Purposes" courses, to direct rejected learners into appropriate general English courses, and to direct the selected students into a suitable combination of "special purpose" and "general proficiency" classes;

- to assess whether a defendant in a court of law could have understood the police-prepared "confession" he had signed and which the police were using in evidence against him;

- to estimate the level of English proficiency in predominantly migrant areas in all States of Australia in order to estimate the English language learning provisions it is necessary to provide;

- to assess the proficiency of Secondary School foreign language learners in a State-wide competency-based approach to school assessment;

- to evaluate the success, strengths and weaknesses of second and foreign language programmes in migrant education centres around Australia and in a tertiary-level foreign languages institute in China;

- to assess the adequacy of foreign students' English proficiency to allow them to undertake academic or vocational studies in education and training institutions in Australia and the United States and to direct them, if necessary, into appropriate remedial programmes.

In brief, the ASLPR is designed to measure practical language proficiency and it can be used for any purpose for which a statement of a learner's practical proficiency is required. It provides that proficiency measurement by indicating where the learner's proficiency falls along the developmental path from zero to native-like. In doing so, it describes second language development and so it also provides a coordinating framework within which programme planning and syllabus design can take place.

III.6 Formal Trials

Formal trials to assess the validity and reliability of the ASLPR have been (or are being) conducted in the context of its use to rate:

(i) adult learners of English as a Second Language,

(ii) adult learners of English as a Foreign Language,

(iii) adolescent learners of English as a Second Language, and

(iv) adult and adolescent learners of French, Italian and Japanese as foreign languages in Australia.

These trials will be reported in full elsewhere [Ingram 1982][(2)] and only a summary of the main results will be given here.

III.6.1 Validity

There are several types of validity which need to be examined in evaluating a new instrument [see Ingram 1982]. Face validity seems to be indicated by the rapidity with which the ASLPR has been accepted around Australia, in a court of law, and by professional registration authorities. Construct and content validity should have been assured by the developmental process in which the scale descriptions drew on psycholinguistic research and were assessed in numerous face-to-face interviews. In assessing concurrent or immediate pragmatic validity, one is essentially asking to what extent the new instrument measures in the same way as established, standardized instruments. This is both a confusing and a practically useful question. It is confusing because existing standardized instruments such as the CELT or TOEFL tests in English or the MLA tests in other languages are knowledge-based, norm-referenced tests whereas the ASLPR is a criterion-referenced test that provides a direct measure of proficiency. Consequently, the standardized tests and the ASLPR are essentially testing different things and a high or stable level of correlation might not be expected. On the other hand, consideration of concurrent validity is practically useful because, if there is a high level of correlation, the ASLPR's measurement of proficiency could be used to explicate in practical skill terms the scores gained on the other tests.

The results on the ASLPR were compared with results on the "Comprehensive English Language Test for Speakers of English as a Second Language", the CELT test [Harris and Palmer 1970]. Table 1 shows the correlation coefficients between the ASLPR ratings and CELT sub-tests for English *Second* Language learners in Australia. The correlations are high, ranging from r = .8 (ASLPR Speaking and CELT Structure) to rho = .96 (ASLPR Reading and CELT Total) with significance levels at or beyond the .001 level. In fact, these coefficients are higher than in many similar studies of the FSI Scale and other standardized tests reported in the literature [e.g., Mullen 1978, Clifford 1978, Carroll 1967, Politzer et al 1982]. However, for English as a *Foreign* Language learners in China, the correlation coefficients were very different. For one group (Table 2), they ranged from rho = .01 (ASLPR Reading and CELT Vocabulary) to rho = .59 (ASLPR Speaking and CELT Structure) with generally moderate significance levels while, for the other group (Table 3), they ranged from rho = -.25 (ASLPR Reading and CELT Listening) to rho = .62 (ASLPR Writing and CELT Listening) with significance levels up to .003 but generally very low.

TABLE 1

Correlations between ASLPR Macroskill Ratings and the Comprehensive English Language Test (CELT) for Adult Learners of ESL in Australia

ASLPR		CELT							
		Structure-A		Listening-A		Vocabulary-A		CELT Total	
		r	rho	r	rho	r	rho	r	rho
Speaking	r/rho	.83	.91	.86	.9	.88	.87	.89	.9
S	n	19	19	18	18	19	19	18	18
	s	.000	.001	.000	.001	.000	.001	.000	.001
Listening	r/rho	.80	.9	.86	.9	.85	.83	.87	.88
L	n	19	19	18	18	19	19	18	18
	s	.000	.001	.000	.001	.000	.001	.000	.001
Writing	r/rho	.88	.94	.89	.9	.91	.91	.93	.95
W	n	19	19	18	18	19	19	18	18
	s	.000	.001	.000	.001	.000	.001	.000	.001
Reading	r/rho	.84	.95	.87	.9	.92	.94	.92	.96
R	n	19	19	18	18	19	19	18	18
	s	.000	.001	.000	.001	.000	.001	.000	.001

TABLE 2

Correlations between ASLPR Macroskill Ratings and the Comprehensive English Language Test (CELT) for Third Year Students of EFL at the Guangzhou Institute of Foreign Languages, Guangzhou, China.

ASLPR		CELT							
		Structure-A		Listening-A		Vocabulary-A		CELT Total	
		r	rho	r	rho	r	rho	r	rho
Speaking	r/rho	.57	.59	.50	.52	.3	.26	.51	.55
S	n	21	21	21	21	21	21	21	21
	s	.003	.002	.01	.008	.094	.13	.009	.005
Listening	r/rho	.52	.55	.50	.52	.32	.27	.51	.58
L	n	21	21	21	21	21	21	21	21
	s	.008	.005	.01	.008	.081	.121	.01	.003
Writing	r/rho	.44	.49	.64	.57	.51	.038	.64	.57
W	n	21	21	21	21	21	21	21	21
	s	.022	.012	.001	.003	.01	.043	.001	.003
Reading	r/rho	.32	.43	.41	.45	.14	.01	.31	.3
R	n	21	21	21	21	21	21	21	21
	s	.079	.025	.031	.021	.274	.483	.083	.096

TABLE 3

Correlations between ASLPR Macroskill Ratings and the Comprehensive English Language Test (CELT) for Second Year Students of EFL at the Guangzhou Institute of Foreign Languages, Guangzhou, China.

ASLPR		CELT							
		Structure-A		Listening-A		Vocabulary-A		CELT Total	
		r	rho	r	rho	r	rho	r	rho
Speaking	r/rho	.1	.22	-.06	.02	.01	.18	.04	.19
S	n	20	20	20	20	20	20	20	20
	s	.33	.18	.4	.46	.5	.23	.44	.21
Listening	r/rho	.38	.47	.16	.2	.4	.47	.39	.5
L	n	20	20	20	20	20	20	20	20
	s	.05	.02	.25	.2	.04	.02	.04	.01
Writing	r/rho	.62	.6	.61	.62	.3	.3	.58	.58
W	n	20	20	20	20	20	20	20	20
	s	.003	.002	.003	.003	.114	.114	.006	.005
Reading	r/rho	.08	.1	-.19	-.25	.13	.07	.03	.002
R	n	20	20	20	20	20	20	20	20
	s	.38	.3	.21	.14	.29	.39	.46	.5

These ambivalent results showing different levels of correlation for different groups of learners are similar to those found in other studies of direct and indirect tests [e.g., Mullen 1978] and seem not to arise from any instability inherent in the ASLPR itself. What they suggest is that, in a *second* language situation where learners have the continual experience of applying their formal knowledge in everyday communication activities, their active communication abilities and passive knowledge in the language will be at similar levels. In a *foreign* language learning situation, especially where traditional grammar-based courses are being followed, there will be a much greater difference between communication skills and passive knowledge with the extent of the relationships depending on each individual's experience, his willingness to "have a go", and his ability to make maximum use of whatever language resources he has for communication purposes. Further evidence for this view was found during studies involving the French version of the ASLPR. Most final year High School students ranged between proficiency levels of 1 and 1+ in their four macroskills. Where levels up to 2 were found (e.g., in a predominantly "migrant" suburb of Sydney), the students were invariably multilinguals who continually had the experience of communicating in their everyday activities in two or more other languages. In other words, though the multilinguals had covered the same content in their French courses as had the other students, they seemed better able to mobilize that content in communication. In all, the correlations found between the ASLPR and the CELT test suggest that where time precludes the use of the ASLPR or where an approximate group measure of *second* language proficiency is required, a test such as CELT can be used. However, much care would need to be exercised in using and interpreting the results of indirect tests with *foreign* language learners if a true picture of practical language proficiency is required.

III.6.2 Reliability

In evaluating an instrument such as the ASLPR, reliability is perhaps the most important question. In examining reliability, two questions were being asked:

1. To what extent can teachers, whether native speakers or non-native speakers, interpret and apply the scale in the same way as do its developers? (inter-rater reliability)

2. To what extent can teachers interpret and apply the scale in the same way, to rate the same learners, on two separate occasions? (intra-rater reliability)

These questions of inter-rater reliability and intra-rater reliability are of fundamental importance for at least two reasons:

1. If a test instrument is to be used by many people, one wants to be sure that they will all use it similarly so that no matter who tests a particular learner the results will be more or less the same.

2. The second reason is even more fundamental and relates to the integrity of the instrument itself. One could assume that if the scale failed to provide a true and coherent picture of how second language proficiency develops, different raters would apply it differently because they focussed on different aspects of the learners' language and matched it with different parts of any description. In other words, low reliability coefficients would suggest that there was something wrong with the assumptions about the path of second language development on which the instrument is based.

To assess reliability, the researchers videotaped learners of each language so that they obtained on videotape and in writing sets of trial material for 16 adult learners of English as a Second Language, 19 adolescent learners of English as a Second Language, and similar sets for adolescent and adult learners of French, Italian and Japanese as Foreign Languages.

Here we shall consider the results involving the adult ESL learners since the other studies are yet to be finalized though procedures are the same and initial results seem to be similar.

The adult learners of ESL were selected to provide about two learners at each defined level of the scale from 0+ to 5. The trial material (videotapes to show the learners' speaking, listening and reading and written scripts to show their writing) were then shown in random order to 21 Australian teachers in Sydney, Melbourne and Adelaide (20 of whom were native-speakers) and to 15 non-native speaking Chinese teachers in Guangzhou, China. (Subsequently the same material has been shown to some 30 American teachers in Los Angeles, though these last results have yet to be analysed.) The teachers were asked to rate each learner on each of the four macroskills of speaking, listening, reading and writing and their ratings were then compared with those assigned by the researchers who developed the scale (Elaine Wylie and the present writer). This process provided a measure of the scale's inter-rater reliability; to obtain a measure of intra-rater reliability, the same material was shown to the same Australian teachers twelve months later and the two sets of ratings were compared.

Tables 4 to 6 show the correlation coefficients when each teacher's ratings for the 16 learners are compared with the researchers' ratings. Clearly, the correlations are at a very high level. This suggests that all the teachers interpreted and applied the scale in essentially the same way as did the researchers who developed it and the Australian teachers interpreted and applied it in essentially the same way on the two different occasions a year apart. Though the levels of correlation with the researchers' ratings are slightly lower for the non-native speaking Chinese teachers, they are still very high, suggesting that even non-native speakers can use the scale reliably.

TABLE 4

Inter-rater Reliability: Australian Teachers and the Developers of the ASLPR

Macroskill	rho		r	
	range	mean	range	mean
Speaking	.91 to .99	.96	.94 to .99	.96
Listening	.89 to .98	.94	.9 to .99	.94
Writing	.93 to .997	.96	.93 to .99	.96
Reading	.91 to .99	.96	.91 to .98	.95

TABLE 5

Inter-rater Reliability: Chinese Teachers and the Developers of the ASLPR

Macroskill	rho		r	
	range	mean	range	mean
Speaking	.88 to .98	.95	.83 to .98	.93
Listening	.88 to .96	.93	.89 to .97	.94
Writing	.88 to .98	.95	.87 to .98	.94
Reading	.88 to .98	.95	.83 to .98	.93

TABLE 6

Intra-rater Reliability: Australian Teachers

Macroskill	rho		r	
	range	mean	range	mean
Speaking	.93 to .99	.97	.93 to .99	.97
Listening	.91 to .98	.96	.92 to .98	.96
Writing	.92 to .98	.96	.91 to .99	.96
Reading	.87 to .98	.96	.91 to .98	.96

rho = Spearman Rank Order Correlation Coefficient
r = Pearson Roduct-Moment Correlation Coefficient

N.B. All correlations were significant at or beyond the .001 level (r) OR .000 (rho)

III.6.3 Trial Conclusions

Although we have seen here only a very brief summary of the results of the formal trialling of the ASLPR, it is sufficient to suggest that the instrument can claim at least acceptable levels of validity and high levels of inter-and intra-rater reliability. In other words, the ASLPR does seem to make valid statements about a learner's language proficiency and users of the ASLPR (whether native or non-native speakers) do seem to be able to interpret and apply it reliably and, therefore, to use it to make reliable assessments of learners' language proficiencies.

IV TESTS IN LANGUAGE PROGRAMMES:

Since the initial release of the earliest version of the ASLPR in February 1979, it has, as we have already seen, been picked up and used very widely around Australia, especially in the ESL field but also in foreign language programmes and the wider community. While the level of this acceptance is gratifying to the ASLPR's developers, nevertheless they have felt some concern on seeing a tendency amongst some teachers, especially in the migrant education programme, to believe that direct testing is the only approach to testing that should be used and the ASLPR the only desirable instrument. There are two principal reasons why this over-use has occurred. First, the emergence of functional approaches to language teaching led some teachers and academics to over-emphasize the ultimate, even though essential, communication goals to the extent that they tended to ignore the elements through which linguistic communication is realized and whose progressive acquisition constitutes partial formative objectives. Second, the functional and communicative approaches to language teaching which rightly gained popularity through the 1970's brought a need for valid and reliable means of assessing practical language skills and so publication of the ASLPR filled an acutely felt need. Nevertheless, the interests of neither the ASLPR and its public acceptability nor successful language teaching and testing are served by the instrument's being used for inappropriate purposes. It is fundamental in all aspects of language teaching that one's purpose must influence one's choice of methods and techniques. Specifically in language testing, one must always ask:

1. What is the purpose of the testing?
2. What is the most appropriate test to achieve this purpose?

Let us consider, therefore but very briefly, some of the purposes for which we test and the instrument or instruments that might be most appropriate to achieve those purposes.

1. To measure how well a learner can use the language, i.e., to measure proficiency:

Fundamentally, this is the purpose for which the ASLPR was developed and it can rightly be used for any purpose to achieve which one requires a measure of practical language skills or proficiency. In fact, this paper has attempted to show that a direct instrument is more appropriate for that purpose, it gives a clearer picture of practical language proficiency, than any other instrument. Furthermore, the finer differentiation amongst low level learners that the ASLPR provides and the fact that it provides a separate measure of each of the four macroskills make it especially useful. The purposes for which measures of general proficiency are required are numerous but, in school, include end-of-course assessment, employment, and tertiary entrance.

2. To give a quick group measure of proficiency:

It cannot be denied that direct instruments are time-consuming to use. Ten to twenty minutes per learner is usually required to assess speaking, listening and reading while the writing test takes more time again but can be done "en masse". There are times when a teacher may want just a quick, approximate idea of how the class as a whole is doing and, for this purpose, an established standardized pencil-and-paper test such as CELT (for English) or one of the MLA tests in other languages may be most appropriate. In addition, although it has not been discussed here, research on the ASLPR and elsewhere has shown that a well constructed cloze test or dictation correlates with proficiency measures in much the same way as does the CELT test with the ASLPR [see Ingram 1982; Oller 1979]. Nevertheless, this paper has already shown that the usefulness of indirect and semi-direct tests in providing even group measures of practical proficiency differs dramatically between second and foreign language learning situations and, probably, between different course types using different methodologies. These differences must, therefore, be taken into account when deciding whether the greater speed of using an indirect or semi-direct test for a group measure warrants the less valid and reliable proficiency measurement provided. Furthermore, it must be remembered that

correlations such as those discussed earlier are *group* comparisons and even a quite high level of correlation as found between the CELT and the ASLPR with second language learners does not mean that all *individuals* are being treated alike. In other words, if one wants to know the proficiency of an individual or to assess what overall progress in developing proficiency he is making, one must use an instrument such as the ASLPR.

3. To measure "special purpose" proficiency:

The ASLPR has been developed to measure general proficiency and has been evaluated in that context. While assessment of general proficiency involves, in part, consideration of register flexibility, general proficiency even amongst native speakers does not ensure proficiency in any particular specialist register. Not all native speakers, for instance, are proficient in the language of civil engineering, neurosurgery, or radiography. If one wants to know if a second language learner is proficient in a specialist register, then one must test in that context. Unfortunately, though a few attempts have been made to develop "special purpose" tests,[3] no well established special purpose instrument is available and the test development task entailed in providing tests in all the "special purpose" areas now being taught would be enormous. It may be that an instrument such as the ASLPR could be readily adapted for that purpose but this has not yet been attempted, the ASLPR has been formally trialled only in the context of general proficiency. How adaptable it would be depends in part on one's concept of "special purpose" proficiency and how it can be developed. If, for example, proficiency in a specialist register is considered to develop over the same parameters as does general proficiency, then the ASLPR could be readily adapted: it would probably require similar "interview" strategies, the general proficiency descriptions would be similar but modified to fit the specialist register, and the "Example" column would have to provide examples of tasks from the specialist register. However, some "special purpose" courses, especially those taught to learners with little or no general proficiency, seem to be based on the view that "special purpose" proficiency develops in quite a different way and emphasis seems to fall not on systematic development but on rote memorization. The effectiveness beyond an elementary level of such an approach must be viewed with scepticism but, if it does work, then it would undoubtedly be necessary to use a different type of test instrument to the ASLPR.

Probably all that one can safely conclude at this stage is that much research remains to be done on "special purpose" testing and, without having formally evaluated the ASLPR or any adaptation of it for this purpose, one must conclude that, although it seems likely that the approach adopted in the ASLPR may be relevant, whether it is or not and how much adaptation would be required remain desirable subjects for future research.

4. To measure achievement in a "graded objectives" course:

The theoretical problem here is not unlike that in testing in a low level "special purpose" course. A "graded objectives" course closely defines the situations, functions and language elements in which the learner is to develop his skills. To the extent that these "graded objectives" contribute to the development of general proficiency, then instruments such as the ASLPR are appropriate to measure that general proficiency. However, the objectives chosen and the teaching approach adopted sometimes seem more like the less creative of the "special purpose" approaches referred to earlier. In such "graded objectives" courses, attainment should be tested more specifically. In particular, rather than varying the situations and tasks in order to gain a picture of the underlying proficiency, as one does in using the ASLPR, in such "graded objectives" courses one would test specifically those language elements and functions in those situations and tasks specified in the course. Exactly what test approaches should be used will differ from course to course. It may include interview techniques or practical task approaches such as used in the ASLPR but it might also include specific knowledge-based approaches and tests of rote-memorized formulae. Again, the general principle applies: consider the objectives, content and methods of the course, consider the purpose of your testing, and select an instrument or test items that will let you achieve the purpose.

5. To measure course attainment (summative purposes):

The observations just made in relation to "graded objectives" courses apply equally to attainment in courses in general. This is sometimes known as "summative" testing. To the extent that the course aims at general proficiency, as most school foreign language programmes do, then the ASLPR is appropriate. If the course aims are different (e.g., to develop knowledge about the language or to develop translation skills), then a different approach (e.g., knowledge-based approaches or

translation) may be necessary and different test types should then be selected (e.g., discrete-point items or a translation test). It may also be that, in considering course attainment, one wants to know to what extent the learner has internalized the content of the course and, if so, then tests directly related to that content may be most appropriate. For example, if one wants to know whether the specified grammatical content has been mastered, then discrete point tests of the grammar taught may give the most direct information. A functional or language task checklist as used on the Lothian progress cards and Waystage Tests may be used to assess achievement in the skills aspects of the course [see Clark 1982a]. If, however, one sees general proficiency as part of the course goals, then the ASLPR should also be used. Certainly, most Secondary School language programmes assume general proficiency as a goal and therefore attainment at least at the main exit points could appropriately be measured using the ASLPR. At intermediate points, it is probably not appropriate to use the ASLPR since general proficiency changes too slowly in unintensive Secondary School programmes for valid and reliable measurements of general proficiency to show much change over a period of a few weeks or even a semester or two. In other words, a combination of approaches is desirable with the ASLPR being used at the exit points and more specific tests related to course content being used in between. Again with course attainment as with other testing, the general principle remains: consider your purpose or purposes and select your test or tests accordingly.

6. To show how well the learners are mastering what is being taught (formative purposes):

This is sometimes called "formative testing" and selection of the appropriate test will depend on exactly what is being taught and what the teacher wants to find out about. If one is interested in whether or not the learner is internalizing the grammar that is being taught, then test items that focus specifically on the grammatical rules will be most appropriate. One could use interview techniques such as are used in applying the ASLPR and simply mark off on a checklist whether or not those grammatical items emerge. However, this is rather haphazard since, even though a skilled interviewer may feel able to so structure a situation that a particular item must be used, in fact the redundancy of language means that there are alternative ways to state most things and so one cannot be sure whether the item failed to emerge

because of the learner's ignorance or simply because he did not need to use it. A discrete point approach or a modified cloze-type where the deleted item lets you focus on the desired grammatical rule would clearly be more efficient and appropriate. If the content includes functions, then a role play approach or even an item such as "Tell me what you would say in a post office to ask the cost of sending a letter to Australia" might be used. To be valid, such test items must, of course, be contextualized in a realistic situation with a contextual purpose for the transaction to occur. The Lothian progress cards and Waystage Tests provide a well-structured way to check the functional and activity content of the course [see Clark 1982a]. Again in formative testing as elsewhere, the general principle remains, consider what you want to find out, consider your purpose, and select your test type and instrument accordingly.

7. To stream learners into an appropriate course:

In the migrant education programme, the ASLPR is the principal basis on which streaming is carried out. To the extent that the different classes are formed on the basis of the learners' general proficiency, it is appropriate to use the ASLPR. However, in principle, other factors ought also to be taken into account since, in streaming, one is aiming to make homogeneous learning groups and factors in addition to proficiency determine how similarly learners will progress: factors such as age, aptitude, family situation, social and psychological distance between the learner and the native speaking community, extent of fossilization or stabilization in the learner's language, attitude to the language, attitude to language learning, motivation, preferred learning strategies, relations between the learner and the teacher or the learner and fellow class members, personal and vocational interests, and so on. In other words, while language proficiency is one factor to be considered in streaming and the ASLPR can be used to measure it, it is only one of many and these other factors may best be measured differently using different test types or even a personal questionnaire.

In addition, one might want to query whether general proficiency is always the most relevant language attainment issue to be considered in streaming. Proficiency relates especially to what Krashen calls "acquired" or informally learned language while streaming is aimed principally at facilitating the teacher's presentation of language to be formally "learned". In other words, if the aim of the streaming is

to place learners in the most appropriate pre-planned courses, then streaming using knowledge-based or discrete function-based tests based on the language to be formally taught might give the most accurate streaming; for this purpose, discrete-point tests or role-play types, for example, might be most appropriate. However, some language programmes, Australia's Adult Migrant Education Program, for example, are planned in relation to general proficiency levels and the dominant developmental trends at each level. In such cases, the ASLPR is an appropriate instrument to be used though, again, only to the extent that language proficiency considerations contribute to decisions on streaming and the other learning factors already referred to must also be taken into account. Other courses, including again the Adult Migrant Education Program, take account in course planning of each learner's own specific needs and additional information is then also needed for course planning purposes on the detail of the learner's language development, personal characteristics and aspirations and, for these, it may well be that the ASLPR has to be supplemented by other instruments. Once again, the general principle remains: for streaming purposes, decide on what basis the classes are to be formed and the streaming done and then decide what are the most appropriate test instruments to use.

8. To diagnose learners' strengths and weaknesses:

In order to choose what instrument to use for diagnostic purposes, clearly one must decide what strengths and weaknesses one wishes to examine. The ASLPR can be used to focus on the learner's practical skills and, by using a checklist to mark off the acceptable and unacceptable forms or deficiencies that occur, one can diagnose the strengths and weaknesses in the elements of the learner's language. However, we have already observed that, for the latter purpose, interviews are rather haphazard, elemental strengths and weaknesses may be missed simply because they do not happen to arise, and test types that isolate specific elements for testing may be preferable. If one wants a comprehensive diagnosis of a learner's language, then a battery of instruments is probably most suitable, e.g., the ASLPR can be used to examine practical proficiency, a standardized discrete-point test might be used to examine syntax and vocabulary, a cloze test can be used to look at discourse, and a cloze and dictation together can be used to randomly sample specific items across the whole expanse of the learners'

language knowledge. The present writer used this battery-approach to examine the nature of the English skills of learners in a foreign language institute in China and more details on this approach can be found in Ingram 1980 and related publications.

9. To evaluate a second language programme:

Programme evaluation raises a large number of contentious and controversial issues. At first sight, one could say that the purpose of a second language programme is to increase a learner's language proficiency, i.e., to take learners from their entry proficiencies toward some point on the way to native-like proficiency. Consequently, it is naively tempting to believe that, to evaluate a programme, all one must do is measure the change in proficiency that occurs from the learners' entry to the course until they exit from it. However, many factors intervene to make a simple measure of proficiency change unsatisfactory as a basis for evaluating the worth of a language programme. These include the learners' aptitudes, their attitudes to the language and its learning, their motivation, their social situation and their perception of how native speakers regard them, the professional competence and language proficiency of the teacher, and so on. In addition, there are other overt and covert goals besides language proficiency that a language programme may seek to achieve (e.g., social confidence) and these must also be taken into account when a programme is being evaluated. In brief, if one is to evaluate a language programme, one is naturally interested in whether or not the learners' language skills improve but one must also take account of the many other overt and covert factors that constitute ancillary goals and objectives or that influence the outcome of language learning. Because of the complexity of the language teaching process (or of education in general for that matter), programme evaluation is better attempted descriptively[4] rather than statistically though the descriptive evaluation should be supported with whatever statistical measurements can be rationally used to inform the description and evaluative judgements. In particular, a diagnostic examination of the learners' language (preferably pre- and post-course) may, as already discussed, provide information on the learners' language with implications for the nature of the programmes followed [e.g., see Ingram 1980].

10. To validate tests:

There are a number of types of validity to be considered when one is developing a new test for wide use. One type, *concurrent* or *immediate pragmatic* validity, is assessed, as we have already observed, by comparing the new test with established tests whose validity has been evaluated and accepted. For the teacher interested in making sure that the test or battery of tests he or she has written or has been using is measuring what he or she wants to measure, it is a useful concept because the new test or battery of tests can be validated by seeing whether the results correlate at a reasonably high level with the results on a possibly more sophisticated standardized test with the same objective. It is important, of course, that the new test or battery of tests is aiming to measure the same thing as the established test against which the correlation is to be made. This, we have already seen, was a major difficulty when the ASLPR was being formally trialled since the CELT is, to a large extent, a test of knowledge rather than proficiency. Where the new test or test battery is aiming to measure proficiency, then it could be validated by correlating the results with the same learners' ratings on the ASLPR. Let us consider an example. If the teacher sees the aim of the course being taught as the development of general proficiency and if, as some education systems require, student evaluation has to be carried out by a process of continual assessment using regular short tests, then the teacher may want to check whether the accumulated results give a reasonably accurate picture of the students' general proficiency. The learners or, if pressure of time is a problem, a cross-section of the learners, can be interviewed and rated on the ASLPR and the results from the teacher's tests and the ASLPR ratings can be compared. Teachers interested in seeing an example of this approach could refer to Ingram 1980 while a very simple and eminently readable explanation of the statistics involved can be found in Crocker 1974.

In this brief consideration of some of the purposes for which one tests, only purposes related to some of the aspects of language proficiency and its contributory elements have been considered. No reference has been made to such factors as motivation, aptitude, attitude, or personality. Clearly instruments such as the ASLPR (or, for that matter, any other test of proficiency or its elements) are not appropriate

for measuring such factors except indirectly to the extent that proficiency changes reflect them. To sum up this brief discussion of tests in language programmes, it is worth reiterating the basic principle: identify the purpose for testing and select the test to match the purpose.

V CONCLUSION:

This paper has outlined the development and trialling of a new direct test, the Australian Second Language Proficiency Ratings. Whereas indirect and semi-direct tests focus on other things, especially linguistic knowledge, the ASLPR, like other direct tests, focusses directly on language proficiency. Formal trials have shown that the ASLPR has a high level of validity and reliability. It was originally developed to measure the proficiency of adults in English as a Second Language but has subsequently been re-considered and is being re-trialled for use with adolescents and adults learning foreign languages. Though the specific foreign languages examined were French, Italian and Japanese and versions of the ASLPR for use with these languages have been developed, there can be no doubt that the ASLPR can appropriately be used wherever one is interested in measuring the general proficiency of adolescents and adults in any language. The choice of the ASLPR or some other instrument should always be seen as dependent on the same basic principle that the test type used and the actual instrument adopted must always accord with the purpose of the testing. Where the purpose is to measure general proficiency, the ASLPR is eminently suitable. Where the purpose is something else, some other instrument may well be more appropriate.

FOOTNOTES

(1) Throughout the development of the ASLPR invaluable assistance has been received from many colleagues including staff of the Education Branch of the Australian Department of Immigration and Ethnic Affairs and of the Language Teaching Branch of the Australian Department of Education, members of the Joint Commonwealth-States Committee on the Adult Migrant Education Program, many individual teachers, and interested colleagues such as Ed Burke, John Mills, Gary Birch, Eila Curtis, Bill Eggington, Kerry Fairbairn, Glyn John, Edwige Coulin, Fiorella Carra, Carlos Zincone, Laura Commins, Roger White, Kyooko Timlock and Ann Beaverson.

(2) Readers interested in seeing the detailed results are invited to write to the author.

(3) See, for example, the Specific Test Packages published by Pergamon Press, Oxford, in "Business", "Science and Technology", and "Everyday International Use".

(4) This is sometimes called an "illuminative approach" to programme evaluation.

REFERENCES

1. Angelis, P.J. 1979. "TOEFL in Recent Years" in Spolsky 1979, pp.101-103.

2. Carroll, John B. 1967. *The Foreign Language Attainments of Language Majors in the Senior Year: A Survey Conducted in U.S. Colleges and Universities.* Cambridge, Mass.: Harvard University Graduate School of Education.

3. Clark, John L. 1982. "Graded Levels of Achievement in Foreign Language Learning (Lothian Region's Project in Modern Languages in Schools)" in *Language Testing Newsletter,* No. 2, Spring 1982, pp.2-4.

4. Clark, John L. 1982a. "How are we to assess language learning in schools?", paper to the Fourth National Languages Conference, Perth, 3rd - 6th September, 1982.

5. Clark, J.L.D. 1972. *Foreign Language Testing: Theory and Practice.* Philadelphia, Pa.: Center for Curriculum Development.

6. Clark, J.L.D. 1978. *Direct Testing of Speaking Proficiency: Theory and Application.* Princeton, N.J.: Educational Testing Service.

7. Crocker, A.C. 1974. *Statistics for the Teacher.* Windsor: N.F.E.R.

8. Foreign Service Institute School of Language Studies 1968. "Absolute Language Proficiency Ratings". Washington, D.C.: U.S. Department of State, mimeograph, reprinted in Clark, 1972, pp.122-123.

9. Harris, D.P. and Palmer, L.A. 1970. *A Comprehensive English Language Test for Speakers of English as a Second Language.* New York: McGraw-Hill.

10. Ingram. D.E. 1980. "Report on the Proficiency Assessment of Second and Third Year Students at the Guangzhou Institute of Foreign Languages, Guangzhou, China, November-December 1980". mimeograph.

11. Ingram, D.E. 1980a. "The Australian Second Language Proficiency Ratings: Their Nature, Development and Trialling", paper to the 1980 RELC Regional Seminar, Singapore, 21st - 25th April, 1980, reprinted in Read 1981, pp.108-136.

12. Ingram, D.E. 1981 "Testing Practical Language Skills", paper to the 1981 FIPLV Congress "Second and Foreign Languages in Education", University of Ibadan, Ibadan, Nigeria, 23rd - 29th August, 1981.

13. Ingram, D.E. 1982. "Report on the Formal Trialling of the Australian Second Language Proficiency Ratings (ASLPR)", Report to the Education Branch, Department of Immigration and Ethnic Affairs, Canberra, August 1982 (mimeograph).

14. Mullen, K.A. 1978. "Determining the Effect of Uncontrolled Sources of Error in a Direct Test of Oral Proficiency and the Capability of the Procedure to Detect Improvement following Classroom Instruction" in Clark 1978, pp.171-189.

15. Oller, John W. 1979. *Language Tests at School.* London: Longman.

16. Oller, J.W. and Spolsky, B. 1979. "The Test of English as a Foreign Language" in Spolsky 1979, pp.92-100.

17. Politzer, Robert L., Shohamy, Elana, McGroarty, Mary. 1982. "Validation of Linguistic and Communicative Oral Language Tests for Spanish-English Bilingual Programs", paper to the pre-conference symposium on language testing, 1982 TESOL Convention, University of Hawaii, Honolulu, Hawaii, 1982 (mimeograph).

18. Read, John A.S. 1981. *Directions in Language Testing*, RELC Anthology Series 9. Singapore University Press.

-oOo-

APPENDIX : KEY HEADINGS FROM

AUSTRALIAN SECOND LANGUAGE PROFICIENCY RATINGS (ASLPR)

S:0 Zero Proficiency	L:0 Zero Proficiency	W:0 Zero Proficiency	R:0 Zero Proficiency
Unable to function in the language.	Unable to comprehend the spoken language.	Unable to function in the written language.	Unable to comprehend the written language.
S:0+ Initial Proficiency	L:0+ Initial Proficiency	W:0+ Initial Proficiency	R:0+ Initial Proficiency
Able to operate only in a very limited capacity within very predictable areas of need.	Able to comprehend only a very restricted range of simple utterances within the most predictable areas of need and only in face-to-face situations with people used to dealing with non-native speakers.	Able to write clearly a limited number of words or short formulae pertinent to the most predictable areas of everyday need.	Able to read only a limited range of essential sight words and short simple sentences whose forms have been memorized in response to immediate needs.
S:1- Elementary Proficiency	L:1- Elementary Proficiency	W:1- Elementary Proficiency	R:1- Elementary Proficiency
Able to satisfy immediate needs using learned utterances.	Able to comprehend readily only utterances which are thoroughly familiar or are predictable within the areas of immediate survival needs.	Able to write with reasonable accuracy short words and brief familiar utterances.	Able to read short simple sentences and short instructions.
S:1 Minimum Survival Proficiency	L:1 Minimum Survival Proficiency	W:1 Minimum Survival Proficiency	R:1 Minimum Survival Proficiency
Able to satisfy basic survival needs and minimum courtesy requirements.	Able to comprehend enough to meet basic survival needs.	Able to satisfy basic survival needs.	Able to read personal and place names, street sians, office or shop designations, numbers, isolated words and phrases, and short sentences.
S:1+ Survival Proficiency	L:1+ Survival Proficiency	W:1+ Survival Proficiency	R:1+ Survival Proficiency
Able to satisfy all survival needs and limited social needs.	Able to satisfy all survival needs and limited social needs.	Able to satisfy all survival needs and limited social needs.	Able to read short texts on subjects related to immediate needs.
S:2 Minimum Social Proficiency	L:2 Minimum Social Proficiency	W:2 Minimum Social Proficiency	R:2 Minimum Social Proficiency
..ble to satisfy routine social demands and limited work requirements.	Able to understand in routine social situations and limited work situations.	Able to satisfy routine social demands and limited work requirements.	Able to read simple prose, in a form equivalent to typescript or printing, on subjects within a familiar context.
S:3 Minimum Vocational Proficiency	L:3 Minimum Vocational Proficiency	W:3 Minimum Vocational Proficiency	R:3 Minimum Vocational Proficiency
Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations on practical, social and vocational topics.	Able to comprehend sufficiently readily to be able to participate effectively in most formal and informal conversations with native speakers on social topics and on those vocational topics relevant to own interests and experience.	Able to write with sufficient accuracy in structures and spelling to meet all social needs and basic work needs.	Able to read standard newspaper items addressed to the general reader, routine correspondence, reports and technical material in his special field, and other everyday materials (e.g. best-selling novels and similar recreational literature).
S:4 Vocational Proficiency	L:4 Vocational Proficiency	W:4 Vocational Proficiency	R:4 Vocational Proficiency
Able to use the language fluently and accurately on all levels normally pertinent to personal, social, academic or vocational needs.	Can comprehend easily and accurately in all personal and social contexts and in all academic or vocational contexts relevant to own experience.	Able to write fluently and accurately on all levels normally pertinent to personal, social, academic or vocational needs.	Able to read all styles and forms of the language pertinent to personal, social, academic or vocational needs.
S:5 Native-like Proficiency	L:5 Native-like Proficiency	W:5 Native-like Proficiency	R:5 Native-like Proficiency
Speaking proficiency equivalent to that of a native speaker of the same socio-cultural variety.	Listening proficiency equivalent to that of a native speaker of the same socio-cultural variety.	Written proficiency equivalent to that of a native speaker of the same socio-cultural variety.	Reading proficiency equivalent to that of a native speaker of the same socio-cultural variety.

NOTE: This table provides only the key headings used in the ASLPR. A copy of the full scale can be obtained by writing to Dr. D.E. Ingram, Department of Languages and Literature, Brisbane College of Advanced Education, Mount Gravatt Campus, Post Office Box 82, Mount Gravatt, Queensland, 4122, Australia.